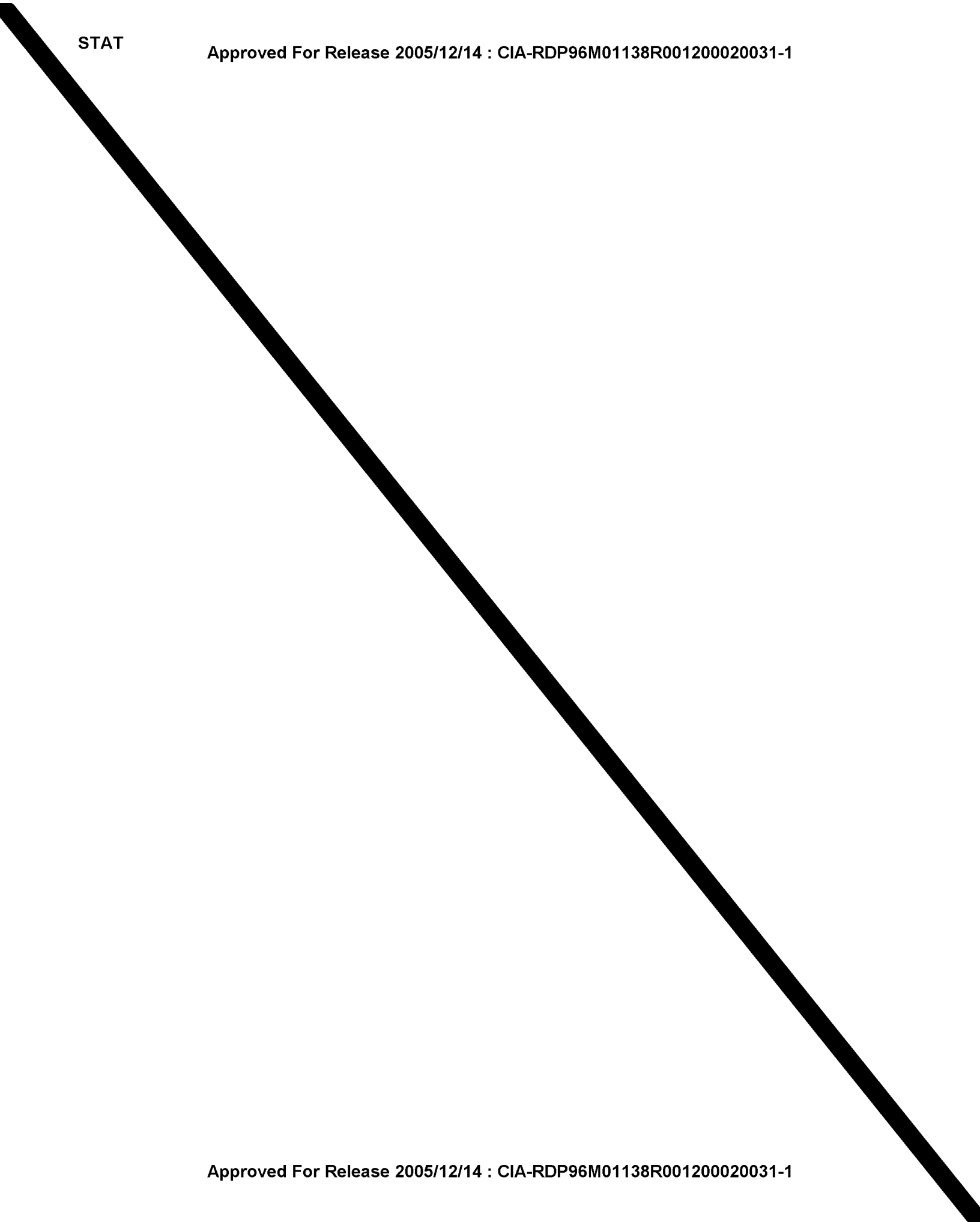


Approved For Release 2005/12/14 : CIA-RDP96M01138R001200020031-1

THE PSYCHOLOGICAL STRESS EVALUATOR: TECHNICAL LIMITATIONS AFFECTING LIE DETECTION

MALCOLM BRENNER
UNIVERSITY OF OREGON

HARVIE BRANSCOMB
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

PREPARED FOR HEARINGS ON S. 1845. SUB-COMMITTEE ON THE CONSTITUTION (COMMITTEE ON THE JUDICIARY), UNITED STATES SENATE, WASHINGTON, D. C., SEPTEMBER 19, 1978.

OUR TESTIMONY CONCERNS TECHNICAL LIMITATIONS IN THE PSYCHOLOGICAL STRESS EVALUATOR (PSE), THE ORIGINAL AND MOST WIDESPREAD OF THE RECENT LIE DETECTION DEVICES WHICH EMPLOY ANALYSIS OF THE HUMAN VOICE. BASED ON OUR RESEARCH EXPERIENCE WITH THIS DEVICE, WE BELIEVE THAT THE PSE IS NOT OF A TECHNICAL QUALITY TO BE USED IN LIE DETECTION AND OUR TESTIMONY DOCUMENTS FIVE TECHNICAL SHORTCOMINGS WHICH AFFECT THE PRESENT INSTRUMENT. THIS EVIDENCE ON TECHNICAL QUALITY IS ESPECIALLY RELEVANT TO LIE DETECTION IN EMPLOYMENT SITUATIONS, SINCE SUCH LIE DETECTION EVIDENCE MAY BE USED AS THE SOLE BASIS OF SERIOUS, UNCONTESTABLE, AND FINAL DECISIONS. EVIDENCE ON TECHNICAL QUALITY IS ALSO RELEVANT TO THE ISSUES OF CONSTITUTIONAL RIGHTS WHICH APPLY TO VOICE LIE DETECTORS BECAUSE OF THE POSSIBILITY OF TESTING SUBJECTS WITHOUT THEIR KNOWLEDGE.

WE DO NOT BELIEVE THAT ALL ASPECTS OF PSE ANALYSIS ARE INVALID. SEVERAL REPORTS PROVIDE EVIDENCE THAT THE PSE MAY BE VALID AS A MEASURE OF PSYCHOLOGICAL STRESS (1,2). THE RATIONALE OF PSE OPERATION (INVOLVING STRESS-SENSITIVE FREQUENCY MODULATION IN THE VOICE) IS CONSISTENT WITH EARLIER ACOUSTICAL EVIDENCE (3). HOWEVER, THERE IS A LARGE DIFFERENCE BETWEEN A PRELIMINARY MEASURE OF

STRESS AND A FINISHED INSTRUMENT WHICH CAN BE APPLIED IN AN AREA AS COMPLEX AS LIE DETECTION. LIE DETECTION ANALYSIS IS CONTROVERSIAL WITH EVEN A HYPOTHETICAL EXCELLENT MEASURE OF STRESS (4), AND WITH A MEASURE SUBJECT TO SERIOUS TECHNICAL LIMITATIONS SUCH ANALYSIS BECOMES EXTREMELY QUESTIONABLE. OUR TESTIMONY CONCERNS SUCH LIMITATIONS IN THE PSE.

THE ORIGINAL RESEARCH REPORTED HERE WAS CARRIED OUT AT HARVARD UNIVERSITY AND THE UNIVERSITY OF OREGON. SPECIFICALLY, IT CONCERNS FIVE TECHNICAL LIMITATIONS:

1) SUBJECTIVITY OF SCORING PSE SCORING IS HIGHLY SUBJECTIVE AND SCORES ASSIGNED TO PARTICULAR PSE PATTERNS DEPEND LARGELY ON THE PARTICULAR JUDGE DOING THE SCORING.

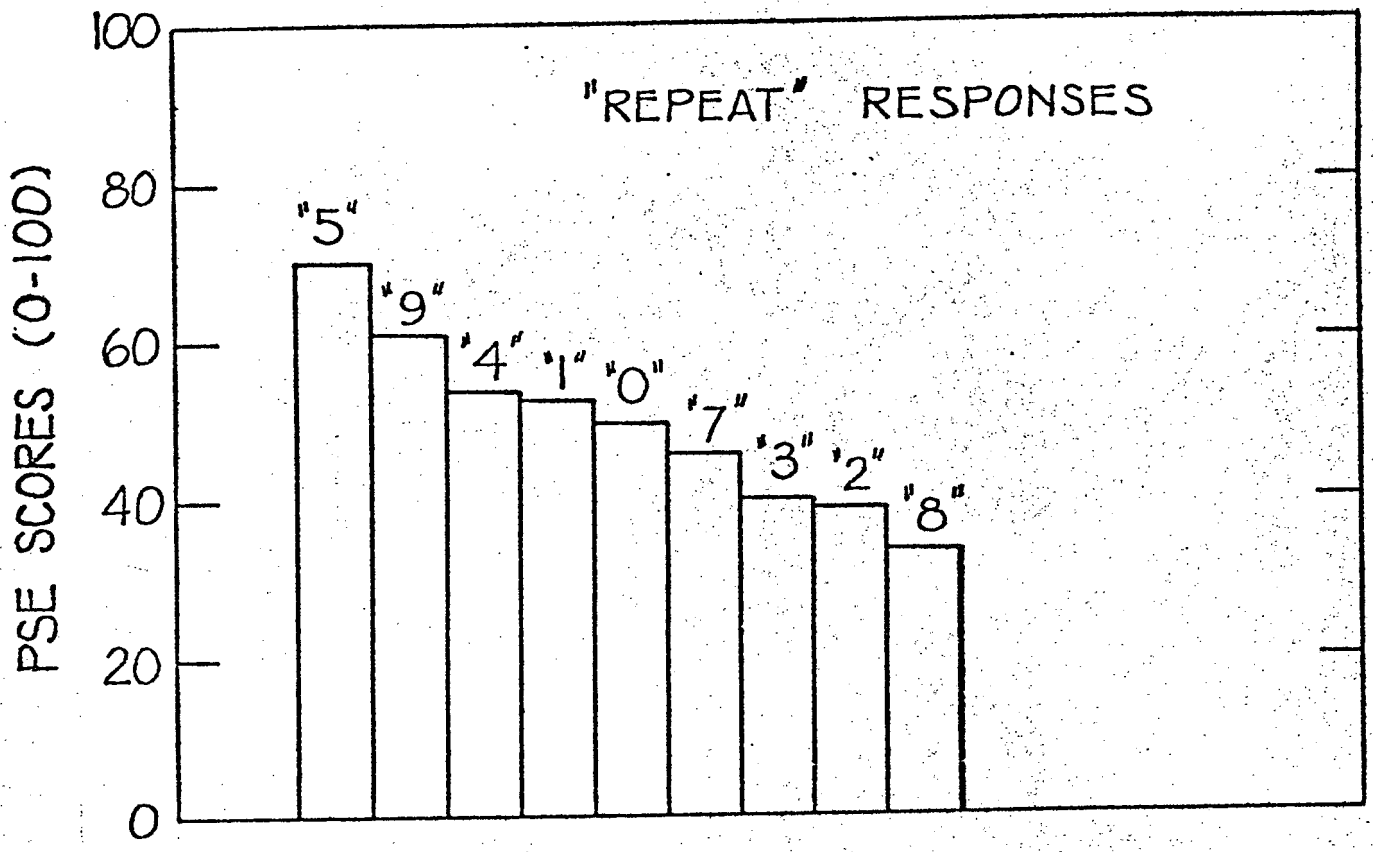
A MATHEMATICAL ESTIMATE OF SCORING SUBJECTIVITY IS AVAILABLE IN THE INTERJUDGE RELIABILITY COEFFICIENT (R), WHICH SUMMARIZES THE DEGREE OF OVERLAP PRESENT IN THE SCORES PROVIDED BY TWO DIFFERENT JUDGES WHO HAVE SCORED THE SAME MATERIAL. THIS RELIABILITY COEFFICIENT RANGES IN VALUE FROM R = .00 TO R = 1.00, WITH THE SIZE OF THE COEFFICIENT DIRECTLY REFLECTING THE DEGREE OF OVERLAP PRESENT. MOST PSYCHOLOGISTS WOULD CONSIDER AN INTER-JUDGE RELIABILITY COEFFICIENT OF R = .80 AS THE MINIMUM REQUIREMENT FOR ANY SERIOUS ASSESSMENT INSTRUMENT.

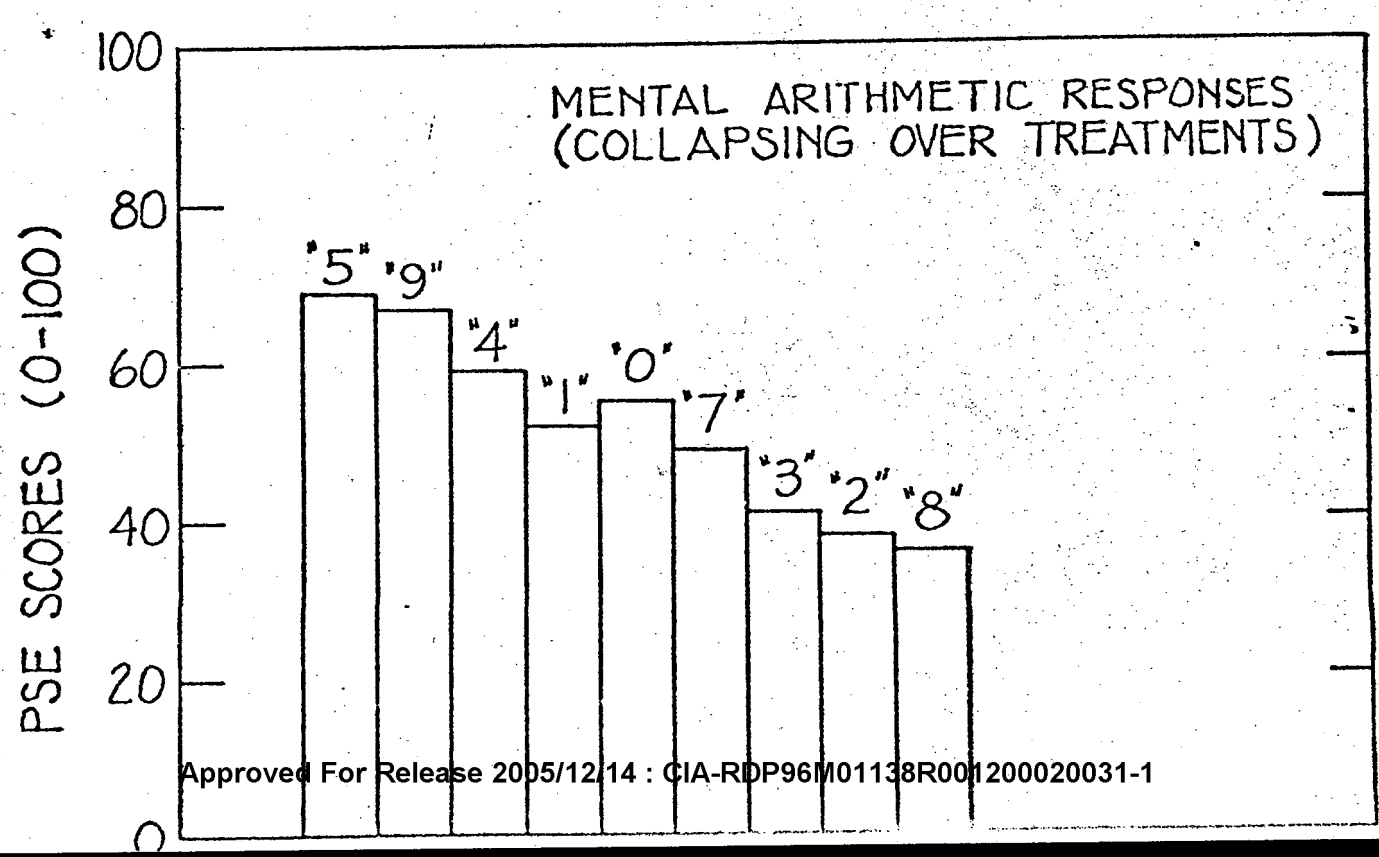
INTERJUDGE RELIABILITY COEFFICIENTS FOR PSE SCORING,

HOWEVER, ARE CONSIDERABLY LOWER THAN R = .80. HORVATH (5) REPORTS A COEFFICIENT VALUE OF R = .38, AND WORTH & LEWIS (6), OF R = .56, FOR MATERIAL DRAWN FROM LABORATORY LIE DETECTION TASKS. OLDER & JENNEY (7) REPORT A COEFFICIENT OF R = .39, AND LEWIS & WORTH (8), OF R = .54, FOR MATERIAL DRAWN FROM TASKS OTHER THAN LIE DETECTION. THESE FOUR COEFFICIENTS, THE ONLY VALUES REPORTED BY OUTSIDE INVESTIGATORS, SUGGEST SERIOUS RELIABILITY PROBLEMS. AN EXAMPLE OF THESE PROBLEMS IS PROVIDED BY A HYPOTHETICAL EXAMPLE OF TWO JUDGES SCORING PSE PATTERNS IN TWO CATEGORIES: HIGH-STRESS OR LOW-STRESS. GIVEN LEWIS & WORTH'S COEFFICIENT VALUE OF R = .56 (THE HIGHEST VALUE REPORTED), THESE JUDGES WOULD BE EXPECTED TO DISAGREE WITH EACH OTHER AT LEAST 22% OF THE TIME (7).

2) RESPONSE WORDS PSE SCORES VARY SYSTEMATICALLY ACCORDING TO THE EXACT WORDS SPOKEN BY THE SUBJECT, AND, PRESUMABLY, THE EXACT LINGUISTIC PROPERTIES OF INDIVIDUAL WORDS.

FIGURE 1 DEMONSTRATES THIS EFFECT, AND SUMMARIZES DATA FOR SIXTEEN SUBJECTS WHO PERFORMED A MENTAL ARITHMETIC EXPERIMENT (3) (1034-1038 SPOKEN RESPONSES ARE SUMMARIZED IN EACH GRAPH). IN THE TOP GRAPH ('"REPEAT" RESPONSES') THE SUBJECTS SIMPLY REPEATED OUT LOUD THE DIGITS FROM "0" TO "9" IN A RANDOM ORDER AS PART OF THE





BASELINE TREATMENT ("6" WAS NOT INCLUDED BECAUSE IT TYPICALLY PROVIDES A PSE PATTERN OF INSUFFICIENT LENGTH TO BE SCORED, A SEVERE EXAMPLE OF RESPONSE WORD DIFFICULTIES). THE DIGITS "5" AND "9" RECEIVED CHARACTERISTICALLY HIGH PSE SCORES, THE DIGIT "8" CHARACTERISTICALLY LOW SCORES, AND THE REMAINING DIGITS INTERMEDIATE SCORES. THIS ROBUST PATTERN APPEARED IN THE DATA OF EVERY SUBJECT TESTED. THIS PATTERN ALSO APPEARS IN THE LOWER GRAPH ("MENTAL ARITHMETIC RESPONSES"), IN A VIRTUALLY IDENTICAL ORDER, DESPITE THE PRESENCE OF A STRONG EXPERIMENTAL MANIPULATION BASED ON THE DIFFICULTY OF MENTAL ARITHMETIC PROBLEMS.[1] IN BOTH GRAPHS, THE PSE SCORING DIFFERENCE BETWEEN HIGH RESPONSE WORDS AND LOW RESPONSE WORDS IS ON THE ORDER OF 2 TO 1.

THE RESPONSE-WORD ARTIFACT SUGGESTS SERIOUS PROBLEMS FOR ANY PSE EXAMINATIONS WHICH USE UNRESTRICTED WORDS OR CONTINUOUS SPEECH. THIS ARTIFACT ALSO HAS DIRECT IMPLICATIONS FOR TRADITIONAL EXAMINATIONS, ESPECIALLY IF IT TURNS OUT THAT "YES" AND "NO" PROVIDE DIFFERENT LEVELS OF PSE-SCORED STRESS.

3) RECORDING QUALITY PSE SCORES TEND TO VARY ACCORDING TO THE QUALITY OF THE AVAILABLE TAPE RECORDINGS.

---

[1] SUBJECTS WERE REQUIRED TO ADD EITHER +4, +3, +1, OR +0 TO EVERY DIGIT IN A STRING OF DIGITS AND REPORT OUT LOUD THE CONVERTED SERIES. THE TIME ALLOWANCE WAS HELD CONSTANT FOR EACH TREATMENT.

EVIDENCE IS PROVIDED BY OLDER & JENNEY (7). THEY PREPARED, UNDER NASA CONTRACT, AN ANALYSIS USING PSE SCORES FOR STRESS CHANGES IN THE VOICES OF SKYLAB ASTRONAUTS AS A FUNCTION OF VARYING WORK LOAD DEMANDS (2040 SPOKEN UTTERNACES WERE INCLUDED IN THE ANALYSIS). THE AVAILABLE TAPE RECORDINGS VARIED CONSIDERABLY IN QUALITY, AND WERE SUBJECTIVELY GROUPED INTO CLASSIFICATIONS OF "GOOD", "FAIR" OR "POOR". ON A 1-5 POINT SCALE USED TO SCORE PSE, OLDER & JENNEY REPORT A DIFFERENCE OF ABOUT 12% IN THE FINAL PSE SCORES AS A DIRECT FUNCTION OF AVAILABLE TAPE QUALITY (PP. 37-39). "GOOD" RECORDINGS SHOWED THE HIGHEST AVERAGE STRESS, AND "POOR" RECORDINGS THE LOWEST.

AN INTERESTING SIDELIGHT OF THE OLDER & JENNEY STUDY IS THE FACT THAT A LARGE SUBSAMPLE OF THE DATA WAS SCORED BY THE CHIEF INSTRUCTOR AT DEKTOR, INC., MANUFACTURERS OF THE PSE (THE INTERJUDGE RELIABILITY COEFFICIENT, AS NOTED ABOVE, WAS R = .39). IRONICALLY, THE CHIEF INSTRUCTOR PROVED TO BE MORE INFLUENCED BY THE ARTIFACT OF TAPE QUALITY THAN THE ROUTINE JUDGE. THE DIFFERENCE IN AVERAGE PSE SCORES WAS 22%.

THIS TAPE-QUALITY ARTIFACT HAS DIRECT RELEVANCE TO INTERROGATION RECORDINGS MADE UNDER FIELD CONDITIONS. THIS ARTIFACT IS ESPECIALLY RELEVANT TO PSE SAMPLES TRANSMITTED OVER THE TELEPHONE, A ROUTINE PROCEDURE

WHICH ALMOST CERTAINLY LOWERS TAPE QUALITY. A JUDGE USING TELEPHONE TRANSMISSION MAY ACTUALLY SCORE A PATTERN WHICH HAS LESS STRESS THAN THE ONE ORIGINALLY PLAYED INTO THE PHONE.

4) TRANSCRIPTION SPEED PSE PATTERNS VARY ACCORDING TO THE SPEED EMPLOYED FOR TRANSCRIBING MATERIAL THROUGH THE DEVICE.

TO DEMONSTRATE THIS EFFECT, WE TRANSCRIBED 217 VOCAL RESPONSES, DRAWN FROM TWO MALE AND ONE FEMALE SUBJECT, AT THE TWO SPEED MOST COMMONLY USED IN PSE ANALYSIS: 1 7/8 IPS AND 15/16 IPS. PSE SCORES DERIVED FROM THE TWO TRANSCRIPTIONS PROVED TO BE EXTREMELY DIFFERENT. CORRELATION COEFFICIENTS BETWEEN THE TWO SCORINGS (COMPUTED THE SAME WAY AS CORRELATION COEFFICIENTS FOR INTERJUDGE RELIABILITY) RANGED BETWEEN R = .43 AND R = .47 FOR EACH SUBJECT. IN 8% OF THE CASES (17 CASES), A PATTERN WHICH SHOWED HIGH PSE STRESS AT ONE SPEED SHOWED LOW PSE STRESS AT A DIFFERENT SPEED.

TECHNICAL INSPECTION SUGGESTS THAT TRANSCRIPTION SPEED DIFFERENCES ARE CAUSED BOTH BY THE SLOW RESPONSIVENESS OF THE TRANSCRIBING PEN AND BY CHANGES IN THE ACTUAL FILTERING CUTOFFS EMPLOYED. WHATEVER THE CAUSE, THIS ARTIFACT DIRECTLY CONTRADICTS COMPANY TRAINING, WHICH LEAVES THE CHOICE OF TRANSCRIPTION SPEED COMPLETELY

AT THE DISCRETION OF THE INTERROGATOR (10).

5) CONSCIOUS CONTROL THE VOCAL RESPONSES MONITORED BY THE PSE MAY BE SUBJECT TO CONSCIOUS CONTROL.

EVIDENCE FOR THIS POSSIBILITY SURFACED IN AN ATTEMPT TO REPLICATE LYKKEN'S GUILTY KNOWLEDGE TASK (3). FIFTEEN MALE AND FIVE FEMALE SUBJECTS UNDERWENT INTERROGATIONS BASED ON ITEMS OF PERSONAL INFORMATION (E. G., "WHAT IS YOUR MOTHER'S FIRST NAME?"). THE SUBJECTS WERE OFFERRED A MONETARY REWARD TO CONCEAL THEIR CORRECT ANSWERS FROM AN INTERROGATOR WHO WOULD SUBSEQUENTLY EMPLOYE PSE ANALYSIS OF THE SUBJECTS' VOCAL RESPONSES. THE INTERROGATOR READ OUT LOUD EACH ITEM FOLLOWED BY SIX POSSIBLE ANSWERS, AND THE SUBJECT REPEATED OUT LOUD ALL POSSIBLE ANSWERS.

LYKKEN REPORTS EVIDENCE FOR THIS TASK USING GALVANIC SKIN RESPONSE (GSR). HE REPORTS THAT THE GSR PROVIDED SIGNIFICANT DETECTION OF GUILTY KNOWLEDGE ITEMS FOR 20 SUBJECTS OUT OF 20 TESTED, EVEN THOUGH SUBJECTS HAD BEEN OFFERRED A MONETARY REWARD AND WERE GIVEN EXTENSIVE PRIOR INFORMATION ON THE INTERROGATION. THE PSE RESULTS FROM OUR EXPERIMENT WERE IN DIRECT CONTRAST: 19 SUBJECTS OUT OF 20 SUCCESSFULLY CONCEALED THEIR CORRECT RESPONSES (FIRST-CHOICE CALLS FOR THESE SUBJECTS, AND DISTRIBUTION OF CALLS, WAS WITHIN THE LEVELS EXPECTED BY CHANCE). ALTHOUGH SEVERAL INTERPRETATIONS

ARE POSSIBLE FOR THIS DIFFERENCE, ONE CLEAR POSSIBILITY IS THAT SUBJECTS ARE ABLE TO VOLUNTARILY INFLUENCE THEIR VOCAL RESPONSES IN A WAY WHICH THEY ARE UNABLE TO INFLUENCE GALVANIC SKIN RESPONSE.

THE ISSUE OF CONSCIOUS CONTROL IS PERHAPS THE MOST SERIOUS ISSUE RAISED IN THIS TESTIMONY, SINCE IT SUGGESTS A BASIC PROBLEM WHICH APPLIES TO ALL VOCAL LIE DETECTORS AND ARGUES FOR A BURDEN OF PROOF ON ALL PERSONS WHO WISH TO SELL VOCAL DEVICES FOR THESE PURPOSES. INCREDIBLY, NONE OF THE OPPONENTS OF VOICE LIE-DETECTORS HAVE RAISED THIS POSSIBILITY, ALTHOUGH THE STRONG CONSCIOUS INFLUENCE ON THE VOICE WOULD SEEM TO MAKE THIS AN OBVIOUS ISSUE OF CONCERN.

THERE IS NOW ENOUGH TECHNICAL EVIDENCE, THEN, TO SERIOUSLY QUESTION THE PSE AS A PRACTICAL LIE DETECTION DEVICE. PROBLEMS OF SCORING SUBJECTIVITY ALONE ARE SUFFICIENTLY SERIOUS IN AVAILABLE LITERATURE TO QUESTION ANY SPECIFIC LEGAL DECISIONS, AND IN PRACTICE THESE SCORING PROBLEMS WOULD BE COMPOUNDED AND MULTIPLIED BY THE EFFECTS OF THE REMAINING ARTIFACTS. THESE TECHNICAL PROBLEMS, IT SHOULD BE NOTED, MAY ALSO APPLY TO THE MORE RECENT MARK II AND HAGOTH LIE DETECTORS WHICH WERE NOT TESTED IN THESE EXPERIMENTS.

THE PRESENCE OF PROBLEMS IN A NEW INSTRUMENT IS NOT SURPRISING, AND SOME OF THE ARTIFACTS DESCRIBED HERE FOR THE PSE ARE TYPICAL FOR ACOUSTICAL MEASURES. WHAT IS SURPRISING IS THE STRENGTH OF THESE EFFECTS IN A DEVICE SOLD FOR A PROCESS AS DELICATE AS LIE DETECTION. LIE DETECTION IS AN EXACTING APPLICATION OF STRESS ANALYSIS, WHICH IS QUESTIONABLE WITH EVEN AN EXCELLENT MEASURE OF STRESS. THE PSE, BY CONTRAST, FAILS TO PASS CERTAIN MINIMAL STANDARDS REQUIRED OF ANY ASSESSMENT MEASURE, AND IS EMPLOYED BY USERS WHO ARE IN AN INAPPROPRIATE POSITION TO RECOGNIZE ITS LIMITATIONS. IT SEEMS INCREDIBLE THAT THIS INSTRUMENT IS PRESENTLY APPLIED IN EMPLOYMENT SITUATIONS, WHERE INDIVIDUALS DO NOT HAVE AN OPPORTUNITY TO QUESTION THE SCIENTIFIC QUALITY OF THE DECISIONS WHICH MAY DIRECTLY AFFECT THEIR EMPLOYMENT.

REFERENCES

1. Borgen, L. A., & Goodman, L. I. Voice analysis of anxiolytic drug effects; preliminary results. Paper presented at meeting of American Society of Clinical Pharmacology and Therapeutics, March, 1976.
2. Brenner, M., Branscomb, H., & Schwartz, G. E. Psychological Stress Evaluator -- two tests of a vocal measure. Unpublished manuscript, 1978 (Submitted for publication).
3. Lieberman, P. Perturbations in vocal pitch. Journal of the Acoustical Society, 33, 1961, 597-603.
4. Lykken, D. T. Psychology and the lie detector industry. American Psychologist, 29, 1974, 725-729.
5. Horvath, F. An experimental comparison of the Psychological Stress Evaluator and the galvanic skin response in detection of deception. Journal of Applied Psychology, 63, 1978, 338-344.
6. Worth, J. W., & Lewis, B. J. An early validation study with the Psychological Stress Evaluator (PSE). Unpublished manuscript, Washington & Lee University (1973).
7. Older, H. J., & Jenney, L. L. Psychological stress measurement through voice output analysis. Planar Corporation, Alexandria, VA, 1975.
8. Lewis, B. J., & Worth, J. W. Transfer of stress through verbal and written communication. Unpublished manuscript, Washington & Lee University, 1974.
9. Wiggins, J. Personality and prediction: Principles of personality assessment. Reading, MA: Addison-Wesley, 1973 (p. 245).
10. Dektor CI/S Inc. PSE-1 Instrument Orientation and Chart Interpretation: a training manual. Springfield, VA (Undated) (p. 4).